Maximum likelihood estimation of regularisation parameters

MAXIMUM LIKELIHOOD ESTIMATION OF REGULARISATION PARAMETERS

Ana Fernandez Vidal, Marcelo Pereyra

School of Mathematical and Computer Sciences, Heriot-Watt University
Edinburgh EH14 4AS United Kingdom

ABSTRACT

This paper presents an empirical Bayesian method to estimate regularisation parameters in imaging inverse problems. The method calibrates regularisation parameters directly from the observed data by maximum marginal likelihood estimation, and is useful for inverse problems that are convex. A main novelty is that maximum likelihood estimation is performed efficiently by using a stochastic proximal gradient algorithm that is driven by two proximal Markov chain Monte Carlo samplers, intimately combining modern optimisation and sampling techniques. The proposed methodology is illustrated with an application to total-variation image deconvolution, where it compares favourably to alternative Bayesian and non-Bayesian approaches from the state of the art.

***Index Terms*—** Image processing, inverse problems, statistical inference, empirical Bayes, stochastic optimisation, Markov chain Monte Carlo methods, proximal algorithms.

1. INTRODUCTION

Image processing problems often require solving a high-dimensional inverse problem that is ill-conditioned or ill-posed. Canonical examples include image deconvolution [1, 2], compressive sensing [3, 4], super-resolution [5, 6], tomographic reconstruction [7, 8], inpainting [9, 10], source separation [11, 12], fusion [13, 14], and phase retrieval [15, 16]. Solving these challenging inverse problems has stimulated significant research in the image processing literature over the past three decades. In particular, convex imaging inverse problems have received a lot of attention in the late, leading to significant advances in models, methods, and algorithms for this class of problems [34].

Most imaging methods to solve inverse problems - whether formulated in a variational or statistical framework [17], or otherwise - use regularisation to make the estimation problem well-posed. In this paper, we focus on the difficult problem of selecting the values of the so-called regularisation parameters that control the amount of regularisation enforced [18, 19].

The remainder of the paper is organised as follows. Section 2 introduces notation and the class of imaging problems considered. Section 3 presents the proposed maximum likelihood estimation method for selecting the values of regularisation parameters. In Section 4, we illustrate the methodology with an application to image deconvolution with a total-variation prior. For comparison, we also present the results obtained with two alternative approaches from the state of the art (the non-Bayesian SUGAR method [20] and the hierarchical Bayesian approach of [19]). Conclusions and perspectives for future work are finally reported in Section 5.

2. PROBLEM STATEMENT

We consider the estimation of an unknown image $\boldsymbol{x} \in \mathbb{R}^n$ from an observation $\boldsymbol{y} \in \mathbb{C}^m$ related to $\boldsymbol{x}$ by a statistical model with likelihood function $p(\boldsymbol{y}|\boldsymbol{x})$ of the form

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \exp\{-g_{\boldsymbol{y}}(\boldsymbol{x})\}\,,$$

with $g_{\boldsymbol{y}}$ convex and Lipschitz continuously differentiable with constant L. This class includes important models, such as Gaussian linear models of the form $g_{\boldsymbol{y}}(\boldsymbol{x}) = \|y - A\boldsymbol{x}\|_2^2/2\sigma^2$, for a linear operator $A \in \mathbb{C}^{m\times n}$ and $\sigma^2 > 0$.

Following a Bayesian approach, we seek to use prior knowledge about $\boldsymbol{x}$ to improve our estimation results. Here we model $\boldsymbol{x}$ as a random vector with prior distribution

$$p(\boldsymbol{x}|\theta) = \exp\{-\theta\,\varphi(\boldsymbol{x})\}/C(\theta)\,,$$

promoting solutions with desired regularity properties. These properties are encoded in φ, which we also assume to be convex, lower semicontinuous and proper, but possibly non-smooth (e.g., analysis priors of the form $\varphi(\boldsymbol{x}) = \|B\boldsymbol{x}\|_1$ with dictionary $B \in \mathbb{R}^{p\times n}$). Finally, for all $\theta \in [0, \infty[$, we assume

$$C(\theta) = \int_{\mathbb{R}^n} \exp\{-\theta\,\varphi(\boldsymbol{x})\}\mathrm{d}\boldsymbol{x} < \infty\,,$$

such that $p(\boldsymbol{x}|\theta)$ is a proper density function. Then, using Bayes' theorem we obtain the posterior distribution [21]

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y},\theta) &= p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\theta)/p(\boldsymbol{y}|\theta)\,, \\ &\propto \exp\{-g_{\boldsymbol{y}}(\boldsymbol{x}) - \theta\,\varphi(\boldsymbol{x})\} \end{aligned} \tag{1}$$

which underpins all inferences about $\boldsymbol{x}$ given $\boldsymbol{y}$. For example, the popular maximum-a-posteriori (MAP) estimator given by

$$\hat{\boldsymbol{x}}_{MAP} \in \underset{\boldsymbol{x}\in\mathbb{R}^n}{\operatorname{argmin}}\; g_{\boldsymbol{y}}(\boldsymbol{x}) + \theta\,\varphi(\boldsymbol{x})\,. \tag{2}$$

In imaging problems, (2) is often computed efficiently by using proximal convex optimisation algorithms involving the gradient $\nabla g_{\boldsymbol{y}}$ and the proximal operator prox_{φ} (see [22]).

Observe that (1) and (2) are parametrised by $\theta \in \mathbb{R}^+$, which acts as a regularisation parameter that controls the balance between observed and prior information. This parameter can significantly impact inferences about $\boldsymbol{x}$, particularly in problems that are ill-posed or ill-conditioned. Unfortunately, identifying good values for θ is notoriously difficult [18, 19].

The Bayesian framework provides two paradigms to select θ automatically: the hierarchical and the empirical [21]. In the hierarchical paradigm we incorporate θ into the model and operate with an augmented posterior $p(\boldsymbol{x}, \theta|\boldsymbol{y})$. This allows removing θ from the model by marginalisation, followed by inference on $\boldsymbol{x}$ with the marginal posterior $p(\boldsymbol{x}|\boldsymbol{y}) = \int_{\mathbb{R}^+} p(\boldsymbol{x}, \theta|\boldsymbol{y})\mathrm{d}\theta$. Alternatively, one can also jointly estimate $(\boldsymbol{x}, \theta)$ from $\boldsymbol{y}$. These two strategies have been successfully applied to imaging models of the form (1) in [19], and we use them as benchmark in our experiments.

The empirical Bayesian paradigm, which we investigate here, estimates θ directly from $\boldsymbol{y}$ by maximum likelihood estimation; i.e., given a set of admissible values Θ, we compute

$$\hat{\theta}_{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, p(\boldsymbol{y}|\theta)\,, \tag{3}$$

where we recall that $p(\boldsymbol{y}|\theta) = \int_{\mathbb{R}^n} p(\boldsymbol{x}, \boldsymbol{y}|\theta)\ \mathrm{d}\boldsymbol{x}$. This is equivalent to Bayesian model selection on a continuous class of models parametrised by θ and selecting the best model-fit-to-data. Given $\hat{\theta}_{MLE}$, we perform inferences with the so-called empirical Bayes posterior $p(\boldsymbol{x}|\boldsymbol{y}, \hat{\theta}_{MLE})$ [21]. Although decision-theoretically suboptimal, empirical Bayesian approaches are more robust to model misspecification and can outperform hierarchical approaches as a result.

Empirical Bayesian approaches generally deliver good results. The challenge, however, is that the marginal likelihood $p(\boldsymbol{y}|\theta)$ is computationally intractable, making the optimisation problem (3) difficult. The main contribution of this paper is to propose a stochastic optimisation scheme to solve (3) efficiently. A main novelty is that the optimisation is driven by proximal Markov chain Monte Carlo (MCMC) samplers [23].

We emphasise at this point that there are several other approaches to select the values of the regularisation parameters in imaging inverse problems that do not rely on the Bayesian framework. For an excellent overview of classical and Bayesian methods please see [24, 18]. More recently, SURE-type methods based on Stein's unbiased risk estimators have received a lot of attention [25, 26, 27, 20]; here we compare with the state-of-the-art method SUGAR [20].

3. PROPOSED EMPIRICAL BAYES METHOD

3.1. Stochastic gradient MCMC algorithm

We now present our method to solve (3). Assume that $\hat{\theta}$ is the unique root of $\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(\boldsymbol{y}|\theta) = 0$ in Θ, such that, for any $\delta > 0$, $\hat{\theta}$ is the unique solution to the fixed-point equation

$$\theta = P_{\Theta}\left(\theta + \delta\frac{\mathrm{d}}{\mathrm{d}\theta}\log\ p(\boldsymbol{y}|\theta)\right), \tag{4}$$

where $P_{\Theta}(\cdot)$ is the projection onto Θ.

If $\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(\boldsymbol{y}|\theta^{(t)})$ was tractable, we could iteratively solve (3) by using the projected gradient algorithm [28]

$$\theta^{(t+1)} = P_{\Theta}\left(\theta^{(t)} + \delta_t\frac{\mathrm{d}}{\mathrm{d}\theta}\log\ p(\boldsymbol{y}|\theta^{(t)})\right), \tag{5}$$

with any non-increasing positive sequence $\{\delta_t\}_{t=1}^{\infty}$ verifying

$$\lim_{t\to\infty}\delta_t = 0, \quad \sum_{t=0}^{\infty}\delta_t = \infty, \quad \sum_{t=0}^{\infty}\delta_t^2 < \infty.$$

However, because (5) is not tractable, here we propose to use a stochastic variant of the projected gradient algorithm based on a noisy estimate of $\frac{\mathrm{d}}{\mathrm{d}\theta}\log p(\boldsymbol{y}|\theta^{(t)})$. To achieve this we express the fixed-point condition (4) in terms of expectation operators. From Fisher's identity we have that [29]

$$\nabla_{\theta}\log p(\boldsymbol{y}|\theta) = \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}\{\nabla_{\theta}\log\ p(\boldsymbol{x}, \boldsymbol{y}|\theta)\}\,,$$

where $\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}$ is the expectation operator w.r.t. $p(\boldsymbol{x}|\boldsymbol{y}, \theta)$. Then, using that $\boldsymbol{y}$ is conditionally independent of θ given $\boldsymbol{x}$, and the identity $\frac{\mathrm{d}}{\mathrm{d}\theta}\log C(\theta) = -\mathrm{E}_{\boldsymbol{x}|\theta}\{\varphi(\boldsymbol{x})\}$ [30], we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log\ p(\boldsymbol{y}|\theta) = \mathrm{E}_{\boldsymbol{x}|\theta}\{\varphi(\boldsymbol{x})\} - \mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}\{\varphi(\boldsymbol{x})\}\,, \tag{6}$$

where we recall that $\mathrm{E}_{\boldsymbol{x}|\theta}$ denotes expectation w.r.t. $p(\boldsymbol{x}|\theta)$. Finally, replacing (6) in (4) leads to the fixed-point equation

$$\theta = P_{\Theta}\left(\theta + \delta\mathrm{E}_{\boldsymbol{x}|\theta}\{\varphi(\boldsymbol{x})\} - \delta\mathrm{E}_{\boldsymbol{x}|\boldsymbol{y},\theta}\{\varphi(\boldsymbol{x})\}\right), \tag{7}$$

that we can solve efficiently by using a stochastic approximation proximal gradient algorithm (SAPG) [31]. Precisely, we construct an SAPG algorithm driven by two Markov kernels $\mathcal{M}_{\theta}$ and $\mathcal{K}_{\theta}$ targeting the posterior $p(\boldsymbol{x}|\boldsymbol{y}, \theta)$ and the prior $p(\boldsymbol{x}|\theta)$ respectively. See Algo. 1 below.

Algorithm 1 Stochastic approx. proximal gradient algorithm

1: Simulate $\boldsymbol{x}^{(t+1)} \sim \mathcal{M}_{\theta^{(t)}}(\boldsymbol{x}|\boldsymbol{y}, \theta^{(t)}, \boldsymbol{x}^{(t)})$ by using (8),
2: Simulate $\boldsymbol{u}^{(t+1)} \sim \mathcal{K}_{\theta^{(t)}}(\boldsymbol{u}|\theta^{(t)}, \boldsymbol{u}^{(t)})$ by using (9),
3: Set $\theta^{(t+1)} = P_{\Theta}\left[\theta^{(t)} + \delta_t\varphi(\boldsymbol{u}^{(t+1)}) - \delta_t\varphi(\boldsymbol{x}^{(t+1)})\right]$.

Given the high dimensionality involved, it is fundamental to carefully choose the kernels $\mathcal{M}_{\theta}$ and $\mathcal{K}_{\theta}$ driving the SAPG. Without loss of generality, here we use the MYULA Markov kernel proposed recently in [32], which is a state-of-the-art proximal Markov chain Monte Carlo (MCMC) method specifically designed for high-dimensional convex problems that are not smooth. These kernels are a variant of the ULA kernels [33, 34], where the non-smooth term φ is replaced by

Table 1. Average mean squared error $\pm$ standard deviation obtained for six images with different algorithms in deblurring with TV prior. Average execution times expressed in minutes.

Method	SNR=20 dB		SNR=30 dB		SNR=40 dB	
	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time	Avg. MSE	Avg. Time
$\theta^*(Oracle)$	22.95 $\pm$ 3.10	–	21.05 $\pm$ 3.19	–	18.76 $\pm$ 3.19	–
SUGAR	24.14 $\pm$3.19	15.74	23.96 $\pm$ 3.26	20.87	23.94$\pm$ 3.27	20.59
Marginalization	24.67 $\pm$ 3.08	17.27	22.39 $\pm$ 3.07	6.31	19.44 $\pm$ 3.26	6.77
Empirical Bayes	23.24 $\pm$ 3.23	43.01	21.16 $\pm$ 3.24	41.50	18.90 $\pm$ 3.39	42.85

its Moreau envelope that is continuously differentiable and whose gradient involves the proximal operator of φ. Precisely, we use the kernels

$$\begin{aligned}\mathcal{M}_\theta: \quad \boldsymbol{x}^{(t+1)} =&(1-\frac{\gamma}{\lambda})\boldsymbol{x}^{(t)} - \gamma\nabla g_{\boldsymbol{y}}\left(\boldsymbol{x}^{(t)}\right)\\ &+\frac{\gamma}{\lambda}\text{prox}_{\varphi}^{\theta\lambda}\left(\boldsymbol{x}^{(t)}\right) + \sqrt{2\gamma}\boldsymbol{z}^{(t+1)},\end{aligned} \tag{8}$$

and

$$\begin{aligned}\mathcal{K}_\theta: \quad \boldsymbol{u}^{(t+1)} =&(1-\frac{\eta}{\lambda})\boldsymbol{u}^{(t)} + \frac{\eta}{\lambda}\text{prox}_{\varphi}^{\theta\lambda}\left(\boldsymbol{u}^{(t)}\right)\\ &+\sqrt{2\eta}\boldsymbol{z}^{(t+1)},\end{aligned} \tag{9}$$

where $\lambda, \gamma, \eta > 0$ are algorithm parameters, $\text{prox}_{\varphi}^{\theta\lambda}$ is the proximal operator of φ with parameter $\theta\lambda$, and $\boldsymbol{z}^{(t+1)} \sim \mathcal{N}(0, \mathbb{I}_n)$; and where for variance reduction we use the same Gaussian random vector to drive $\mathcal{M}_\theta$ and $\mathcal{K}_\theta$.

Note that these kernels target regularised approximations of $p(\boldsymbol{x}|\boldsymbol{y},\theta)$ and $p(\boldsymbol{x}|\theta)$ improving the convergence speed of the algorithm (see [32] for details) at the expense of introducing an approximation error. As a result, Algo. 1 will have some asymptotic estimation bias. This error is controlled by λ, γ, and δ, and can be made arbitrarily small at the expense of additional computing time (see [32] for details). The bias can also be completely removed by combining (8)-(9) with Metropolis-Hastings steps, as discussed in detail in [23]. However, this may significantly deteriorate convergence speed [32]. A theoretical analysis of the convergence properties of Algo. 1 is currently under investigation and will be detailed elsewhere.

3.2. Implementation guidelines

We now provide some practical guidelines for implementing Algo. 1. Recalling that $\nabla g_{\boldsymbol{y}}$ is Lipschitz continuous with constant $L > 0$, we recommend implementing $\mathcal{M}_\theta$ with $\lambda = 1/L$ and $\gamma \in (\lambda/2(L\lambda+1), \lambda/(L\lambda+1))$. For $\mathcal{K}_\theta$, we use $\lambda \in [5, 10]$ and $\gamma \in (\lambda/2, \lambda)$. Regarding the choice of the sequence $\{\delta_t\}_{t=1}^{\infty}$, a standard choice is $\delta_t = \alpha\, t^{-\beta}/n$ for some $\alpha > 0$ and $\beta \in [0.6, 0.9]$. We emphasise that these recommendations do not seek to define optimal values for specific models, but rather to provide general rules that are simple and robust. For more details please see [32].

4. NUMERICAL EXPERIMENTS

In this section we illustrate the proposed methodology with an application to image deblurring using a total-variation prior. For comparison, we also report the results obtained with SUGAR [20], marginal MAP estimation [19], and by using the optimal or oracle value θ^* that minimises the estimation mean squared error (MSE).

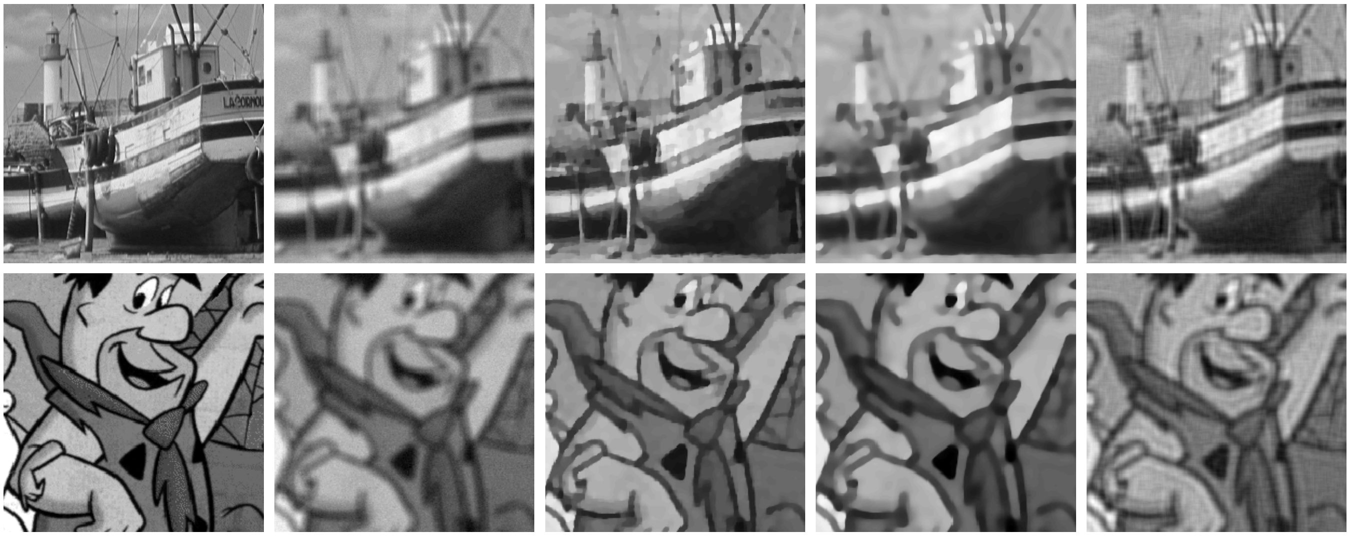
In non-blind image deblurring, the aim is to recover an unknown image $\boldsymbol{x} \in \mathbb{R}^n$ from a blurred and noisy observation $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{w}$, where $\mathcal{A}$ is a circulant blurring matrix, and $\boldsymbol{w} \sim \mathcal{N}(0, \sigma^2\mathbb{I}_m)$. In our experiments, x and y are of size $n = m = 512 \times 512$ pixels, A implements a known uniform blur of size 9×9 pixels, and σ^2 is chosen such that the blurred signal-to-noise-ratio (SNR) is 20 dB, 30 dB, or 40 dB. We use the model (1) with $g_{\boldsymbol{y}}(\boldsymbol{x}) = \|\boldsymbol{y} - \mathcal{A}(\boldsymbol{x})\|_2^2/2\sigma^2$ and $\varphi(\boldsymbol{x}) = TV(\boldsymbol{x})$ (the isotropic total-variation pseudo-norm). We perform all experiments on six standard test images (`barbara, boat, flintstones, lena, man and mandrill`).

For each image, noise level, and method, we compute the MAP estimator $\hat{\boldsymbol{x}}_{MAP}$ (given by (2) or the marginalised MAP in the case of [19]). We then assess the resulting performance by measuring MSE w.r.t. ground truth and computing time. These results are reported in Table 1[1].

We observe from Table 1 that the proposed method performs close to the oracle performance, generally outperforming the other approaches from the state of the art, at the expense of longer computing times. In particular, the proposed method performs remarkably at low and medium SNR values (i.e., 20 dB and 30 dB), and at high SNR values (40 dB) it performs similarly to the marginal MAP method [19]. SUGAR performs less well in these experiments because it minimises a surrogate of the MSE that degrades in problems that are ill-posed or ill-conditioned.

For illustration, Fig. 1 shows the results obtained for two of the test images and by using each of the methods considered (the displayed images correspond to the 20 dB SNR setup). We display a close-up on an image region containing fine detail and sharp edges. In Fig. 2 we provide further details for the `boat` experiment where we show how $\theta^{(t)}$

[1]All experiments were conducted on an Intel i7-7700@3.60GHz workstation running Matlab R2017b.



(a) Original (b) Degraded (c) Emp. Bayes (d) Marginalization (e) SUGAR

Fig. 1. Deblurring with TV prior - Close-up on Boat and Flinstones test images: (a) Underlying image $\boldsymbol{x}$, (b) blurred and noisy (SNR=20 dB) image $\boldsymbol{y}$, (c)-(e) MAP estimators obtained through Empirical Bayes, marginalization and SUGAR methods respectively.

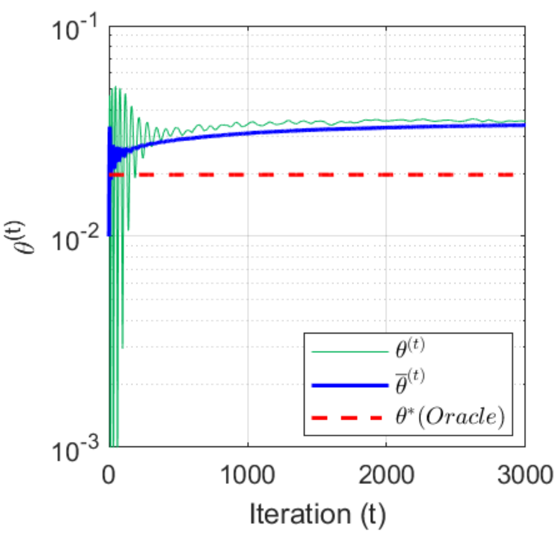


Fig. 2. Deblurring with TV prior experiment with `boat`: evolution of the iterates $\theta^{(t)}$ and $\overline{\theta}^{(t)}$ for the proposed method. The Oracle's θ^{*} is plotted as a reference.

progresses throughout the iterations. We also define the cumulative mean $\theta^{(t)}$ given by

$$\overline{\theta}^{(t)} = (1/t) \sum_{j=1}^{t} \theta^{(j)}.$$

Finally, note that in all experiments we implemented the proposed method using $\Theta = [0.001, 1]$, initial condition $\theta^{(0)} = 0.01$, sequence of step-sizes $\delta_t = 0.1\ t^{-0.8}/n$, and by warm-starting Algo. 1 with 200 burn-in iterations with fixed $\theta^{(t)} = \theta^{(0)}$, followed by $2\,000$ iterations to compute $\hat{\theta}$. For kernel (8) we used $\lambda = 2$ and selected γ based on the Lipschitz constant L (which depends on σ^2), leading to the values $\gamma_{20} = 1.82$, $\gamma_{30} = 1$ and $\gamma_{40} = 0.182$ for SNR values 20 dB, 30 dB and 40 dB respectively. For kernel (9) we used $\lambda = \eta = 5$ in all experiments. Also, because the prior associated with the total-variation norm is not proper (i.e., $\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \infty$), we implemented (9) by using $\varphi(\boldsymbol{x}) = TV(\boldsymbol{x}) + \kappa\|\boldsymbol{x}\|_2^2$ with a small value κ and removed this correction by importance sampling in Step 3 of Algo. 1.

5. CONCLUSIONS

This paper presented an empirical Bayesian method to estimate regularisation parameters directly from observed data in imaging inverse problems that are convex. The method proceeds by maximum marginal likelihood estimation. This marginal likelihood is computationally intractable, and we addressed this difficulty by proposing a carefully designed stochastic optimisation algorithm based on a stochastic approximation scheme driven by two proximal MCMC kernels. The proposed methodology was illustrated with an application to non-blind image deconvolution, where it achieved close-to-optimal performance and outperformed approaches from the state of the art in terms of estimation MSE. A detailed theoretical analysis of the convergence properties of the proposed methodology is currently under investigation.

Future work will focus on extending the methodology to problems involving several unknown regularisation parameters, and on reducing computing times by accelerating the Markov kernels driving the stochastic approximation algorithm.

6. REFERENCES

[1] Silvia Bonettini, Anastasia Cornelio, and Marco Prato, "A new semi-blind deconvolution approach for fourier-based image restoration: an application in astronomy," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1736–1757, 2013.

[2] Bruno Amizic, Leonidas Spinoulas, Rafael Molina, and Aggelos K Katsaggelos, "Compressive blind image deconvolution," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3994–4006, 2013.

[3] David L Donoho, "Compressed sensing," *IEEE Transactions on information theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[4] Emmanuel J Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[5] S Derin Babacan, Rafael Molina, and Aggelos K Katsaggelos, "Variational bayesian super resolution," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 984–999, 2011.

[6] Veniamin I Morgenshtern and Emmanuel J Candes, "Super-resolution of positive sources: The discrete setup," *SIAM Journal on Imaging Sciences*, vol. 9, no. 1, pp. 412–444, 2016.

[7] José M Bioucas-Dias and Mário AT Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *IEEE Transactions on Image processing*, vol. 16, no. 12, pp. 2992–3004, 2007.

[8] Michael Lustig, David Donoho, and John M Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[9] Raymond H Chan, Junfeng Yang, and Xiaoming Yuan, "Alternating direction method for image inpainting in wavelet domains," *SIAM Journal on Imaging Sciences*, vol. 4, no. 3, pp. 807–826, 2011.

[10] Julien Mairal, Michael Elad, and Guillermo Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on image processing*, vol. 17, no. 1, pp. 53–69, 2008.

[11] José MP Nascimento and José MB Dias, "Vertex component analysis: A fast algorithm to unmix hyperspectral data," *IEEE transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 898–910, 2005.

[12] Zhengming Xing, Mingyuan Zhou, Alexey Castrodad, Guillermo Sapiro, and Lawrence Carin, "Dictionary learning for noisy and incomplete hyperspectral images," *SIAM Journal on Imaging Sciences*, vol. 5, no. 1, pp. 33–56, 2012.

[13] Miguel Simões, José Bioucas-Dias, Luis B Almeida, and Jocelyn Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2015.

[14] Gloria Haro, Antoni Buades, and Jean-Michel Morel, "Photographing paintings by image fusion," *SIAM Journal on Imaging Sciences*, vol. 5, no. 3, pp. 1055–1087, 2012.

[15] Emmanuel J Candes, Yonina C Eldar, Thomas Strohmer, and Vladislav Voroninski, "Phase retrieval via matrix completion," *SIAM review*, vol. 57, no. 2, pp. 225–251, 2015.

[16] Jos M Bioucas-Dias and Gonalo Valadao, "Phase unwrapping via graph cuts," *IEEE Transactions on Image processing*, vol. 16, no. 3, pp. 698–709, 2007.

[17] Jari Kaipio and Erkki Somersalo, *Statistical and computational inverse problems*, vol. 160, Springer Science & Business Media, 2006.

[18] Rafael Molina, Aggelos K Katsaggelos, and Javier Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 231–246, 1999.

[19] Marcelo Pereyra, José M Bioucas-Dias, and Mário AT Figueiredo, "Maximum-a-posteriori estimation with unknown regularisation parameters," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 230–234.

[20] Charles-Alban Deledalle, Samuel Vaiter, Jalal Fadili, and Gabriel Peyré, "Stein unbiased gradient estimator of the risk (sugar) for multiple parameter selection," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 2448–2487, 2014.

[21] Christian Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, Springer Science & Business Media, 2007.

[22] Patrick L Combettes and Jean-Christophe Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, pp. 185–212. Springer, 2011.

[23] Marcelo Pereyra, "Proximal markov chain monte carlo algorithms," *Statistics and Computing*, vol. 26, no. 4, pp. 745–760, 2016.

[24] Alan M Thompson, John C Brown, Jim W Kay, and D Michael Titterington, "A study of methods of choosing the smoothing parameter in image restoration by regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 326–339, 1991.

[25] Charles M Stein, "Estimation of the mean of a multivariate normal distribution," *The annals of Statistics*, pp. 1135–1151, 1981.

[26] Sathish Ramani, Thierry Blu, and Michael Unser, "Monte-carlo sure: A black-box optimization of regularization parameters for general denoising algorithms," *IEEE Transactions on Image Processing*, vol. 17, no. 9, pp. 1540–1554, 2008.

[27] Jean-Christophe Pesquet, Amel Benazza-Benyahia, and Caroline Chaux, "A sure approach for digital signal/image deconvolution problems," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4616–4632, 2009.

[28] Yves F Atchadé, Gersende Fort, and Eric Moulines, "On perturbed proximal gradient algorithms," *J. Mach. Learn. Res*, vol. 18, no. 1, pp. 310–342, 2017.

[29] Yves F. Atchadé, "A computational framework for empirical Bayes inference," *Statistics and Computing*, vol. 21, no. 4, pp. 463–473, 2011.

[30] Marcelo Pereyra, Nicolas Dobigeon, Hadj Batatia, and Jean-Yves Tourneret, "Computing the cramer–rao bound of markov random field parameters: application to the ising and the potts models," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 47–50, 2014.

[31] Gersende Fort, Edouard Ollier, and Adeline Samson, "Stochastic proximal gradient algorithms for penalized mixed models," *arXiv preprint arXiv:1704.08891*, 2017.

[32] Alain Durmus, Eric Moulines, and Marcelo Pereyra, "Efficient bayesian computation by proximal markov chain monte carlo: when langevin meets moreau," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 473–506, 2018.

[33] Peter J Green, Krzysztof Latuszyński, Marcelo Pereyra, and Christian P Robert, "Bayesian computation: a summary of the current state, and samples backwards and forwards," *Statistics and Computing*, vol. 25, no. 4, pp. 835–862, 2015.

[34] Marcelo Pereyra, Philip Schniter, Emilie Chouzenoux, Jean-Christophe Pesquet, Jean-Yves Tourneret, Alfred O Hero, and Steve McLaughlin, "A survey of stochastic simulation and optimization methods in signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 224–241, 2016.